

Contents

1	Introduction	5
1.1	Information Extraction	6
2	State of the Art	6
2.1	Information Extraction from HTML Documents	7
2.1.1	Wrappers	8
2.2	Document Code Modeling	8
2.3	Wrapper Induction Approaches	9
2.3.1	Methods Based on the Grammatical Inference	9
2.3.2	Methods Based on Relational Learning	10
2.4	HTML Code Analysis	11
2.5	Conceptual Modeling	11
2.6	Semi-automatic Wrapper Construction	11
2.7	Logical Document Structure	11
2.7.1	Visual Analysis of HTML Documents	12
2.8	Logical Document Discovery	13
3	Motivation and Goals of the Thesis	14
4	Visual Modeling Approach	16
4.1	Visual Information Analysis	16
4.1.1	Page Layout Model	17
4.1.2	Text Attribute Model	18
4.2	Logical Document Structure	20
4.3	Information Extraction from the Logical Structure	21
4.4	Logical Document Discovery	22
5	Experimental Results	23
6	Summary of Contributions	23
7	Conclusions	25
	References	25
	Author's Curriculum Vitae	29